DOCUMENT RESUME

ED 262 085 TM 850 566

AUTHOR Cope, Ronald T.
TITLE Use versus Nonuse of Repeater Examinees in Common Item Linear Equating with Nonrandom Groups.
PUB DATE Jun 85
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association, 69th, Chicago, IL, March 30-April 4, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; *Equated Scores; Measurement Techniques; Postsecondary Education; Professional Continuing Education; *Statistical Analysis; *Testing Problems; *Testing Programs
IDENTIFIERS Anchor Test Study; *Linear Equating Method; *Repeated Testing

ABSTRACT

This study considers the use of repeaters when test equating. The subjects consist of five groups of applicants to a professional certification program. Each group comprises first time examinees and repeaters. The procedures include a common item linear equating with nonrandom groups, use of equating chains, and the use of total examinee group versus first-timers only. The failure of the results to show uniformly and substantially more accurate equating under use of only first-time (nonrepeater) examinees casts doubt upon the practice of routinely excluding repeater examinees in deriving equating functions for this testing program. The overall pattern of results does appear to favor slightly the use of first-timers. But the relative accuracy of equating with first-timers versus total group also appears to depend partly on the specific equating method (Tucker Equally Reliable versus Levine Equally Reliable). Further research is needed to determine whether eliminating larger proportions of repeaters produces differences greater than those found. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED262085

Use Versus Nonuse of Repeater Examinees in Common Item Linear Equating with Nonrandom Groups

Ronald T. Cope

The American College Testing Program

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Cope, Ronald T.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Presented at the Annual Meeting of the

American Educational Research Association

Chicago, April 1985

Revised June 1985

TM 850 566

THE PROBLEM

Two common rules of thumb in test equating are (1) Use as many examinees as possible, and (2) Exclude repeaters--examinees who have taken the test before.

These two rules conflict. Rule 1 tells us on the one hand to obtain the most stable statistics and equating functions by including all available examinees. Rule 2, on the other hand, tells us to exclude some examinees from our calculations, and implies that something is wrong with using repeaters.

Why exclude repeaters? The reason is that we expect the usually self-selected repeaters to differ in various ways from first-time examinees. On some tests--certification tests, for example--we expect repeaters to score lower on the average than first-timers. If repeaters' scores tend to accumulate close to the bottom of the score scale, we ought to suspect serious violation of the assumptions underlying our equating technique (Angoff, 1971, pp. 580-581; Braun & Holland, 1982, pp. 25-27). Repeaters may differ from first-timers in motivation and in their profile of the abilities the test measures. These differences might distort the equating relationship.

Also, repeaters might have seen some or all of the common equating items when they took previous forms. Suppose that we include repeaters in our current or "new" form equating group even though we excluded them from the previous-form or "old" equating group. We might expect the repeaters' previous exposure to common items to produce the following effect: Repeaters perform better on the common items than they otherwise would have, and the average common items score on the new form is thereby spuriously high. Consequently, the new form examinee group appears to be more able than it really is. Previously unused noncommon items appear to be harder than they really are in relation to the common items, and so the entire new form appears

harder than it really is. As a result, old form raw scores convert to erroneously low raw scores on the new form, and a given raw score on the new form yields an erroneously high scaled score.

Finally, repeaters have had the benefit of practice with an entire preceding test form (which may or may not have items in common with the new form). Practice effects might further distort the equating. Note that the Angoff II equating design (Angoff, 1971, pp. 573-576) requires each examinee to take both forms to be equated; this design requires all examinees to be repeaters. Counterbalancing the order of administration of the forms must be used to manage exposure effects. But we cannot counterbalance in a testing program that uses a single test form at one administration and another form at a later administration.

These considerations give reason to be careful about using repeaters for test equating. What actually does happen when repeaters are included? Do the resulting equating functions differ from those obtained when repeaters are excluded? The purpose of this study is to try to answer these questions.

METHOD

Subjects

The subjects of this study consist of five groups of applicants to a professional certification program. Each of these five groups comprises first time examinees, repeaters, and a small proportion of examinees who neglected to indicate whether they had taken the test before. Each one of these five examinee groups took one of five forms of the certification test in a particular administration. Only one form of the test is used per administration and the form is not reused.

Table 1 gives the breakdown of examinee groups according to repeater status. The letter designations indicate the chronological order of administration of the forms: Form A was administered first of the five forms, Form B was administered after Form A, etc.

Insert Table 1 about here

The five groups do not represent completely exclusive sets of examinees. For example, a repeater on Form C might have taken Form A, Form B, both of them, or one or more forms that preceded Form A. Analysis of a recent subgroup of repeaters shows that nearly half of them have taken the test more than twice.

Measures

Measures are five forms of a multiple-choice professional certification test with a raw score range of 0 to 280. The test is administered twice a year at six-month intervals. For each administration a new form is constructed, used in the administration, and then retired. Except for retired items used in a sample test, no future, current, or past test items are released to examinees or otherwise publicized.

Procedure

I. Equating Design. In this particular certification test program, common item linear equating with nonrandom groups is used under the assumption of equally reliable forms (estimated reliabilities range from .94 to .95). Under use of an internal anchor test, this equating design requires that a given form share a set of items--called link items, equators, or the anchor

test--with preceding forms to which it is equated. These link items should accurately represent the item statistics and content specifications of the test forms (Petersen, Marco, & Stewart, 1982; Klein & Jarjoura, in press).

II. Use of Equating Chains. To compare the accuracy of equating under conditions of inclusion versus exclusion of repeater examinees, I identified three partly-overlapping chains of equating links among the five test forms (see Petersen, Cook, & Stocking, 1983; Klein & Jarjoura, in press). Figure 1 depicts these equating chains. An arrow from one form to another indicates both an equating between the forms and the direction of the equating. For example, in Chain 1 the arrow from Form B to Form A indicates that (1) Form B has been equated to Form A, and (2) Form B is the "new" or more recent form, and Form A is the "old" or earlier form. When we proceed to the Form B → Form C link of Chain 1, Form B becomes the old form and Form C is the new form.

Insert Figure 1 about here

These chains require reversing the normal or chronological equating order for one pair of forms in each chain. This is true for Form A → Form C of Chain 1, Form A → Form D of Chain 2, and Form B → Form E of Chain 3. For example, in equating Form A to Form C, Form A (the earlier) would ordinarily be the old form; Form C (administered a year later than Form A) would be the new form equated back to Form A.

Chains such as these help greatly in comparing the accuracy of different equating procedures. They allow us to refer to a single score scale--raw scores of a single form of a chain--and to equate this form through the chain back to itself. I call this form the "base form." In Figure 1 the base forms (underlined) are Form A in Chain 1, Form A in Chain 2, and Form B in Chain 3.

To equate the base form to itself, proceed as follows: Consider Chain 1 with base Form A. Begin with Form A as the "new" form and equate it to "old" Form C (note again that in this link we reverse the actual chronological order). Obtain the linear equating function $C = m_1A + b_1$ relating raw scores on Form A to raw scores on Form C. In this equation, C is the Form C raw score, m_1 is the slope of the equating function, A is the Form A raw score, and b_1 is the C-intercept. Next Form C becomes the new form, and we equate it to old Form B. Obtain a second equating function $B = m_2C + b_2$. Next equate Form B as the new form to old Form A. Obtain a third equating function $A' = m_3B + b_3$.

Now take the composite of these three equating functions to obtain

$$A' = m_3(m_2[m_1A + b_1] + b_2) + b_3 \quad ,$$

$$A' = m_1m_2m_3A + m_2m_3b_1 + m_3b_2 + b_3 \quad ,$$

$$A' = m_4A + b_4 \quad . \qquad (1)$$

Equation 1, the "final equating function," relates given raw scores, A, on Form A as "new" test around the chain to raw scores, A', on Form A as "old" test. If there were no error introduced by the three equatings, the slope, m_4, and intercept, b_4, of Equation 1 would be, respectively, 1 and 0. A given raw score would cycle around the chain to its precise equivalent. However, test equating is susceptible to various kinds of error; we therefore expect the slope and intercept to differ from 1 and 0.

III. Use of Total Examinee Group Versus First-Timers Only. The base form of each of the three equating chains was equated to itself under the following two procedures:

1. Use the total examinee group for each form of the chain to obtain statistics leading to the equating functions; the total group includes first-time examinees ("first-timers"), repeaters, and those who did not indicate their repeater status, and
2. Use only the subgroup of first-timers to obtain statistics yielding the equating functions.

For each of these two procedures, two different equating methods were used: Tucker Equally Reliable and Levine Equally Reliable, for brevity referred to here as the Tucker and the Levine methods.

RESULTS

Tables 2-4 give the results of equating the base forms to themselves through the three equating chains. Over most of the possible range of raw scores the first-timer examinee group produces slightly more accurate results than the total examinee group under use of the Tucker method. However, under use of the Levine method the results do not decisively favor use of one group over the other. Using the first-timer group in Chain 3 under the Levine method yields more accurate equivalents throughout the entire range of scores. But in Chain 1 and Chain 2, first-timer and total group equivalents contend for being the more accurate as one traverses the range of possible scores.

Insert Tables 2, 3, and 4 about here

To facilitate comparisons of results under the use of the different groups, methods, and equating chains, I have computed estimates of (1) mean equating error or bias and (2) root-mean-squared error of equating (see Klein & Jarjoura, in press).

Estimated mean equating error or bias is defined as

$$\widehat{\text{BIAS}} = \bar{X}' - \bar{X} \quad ,$$

in which $\bar{X}$ is the mean base form raw score, and $\bar{X}'$ is the mean of the raw score equivalents obtained by equating the base form to itself through a chain.

Estimated root-mean-squared error, $\widehat{\text{RMSE}}$, is defined by the expression

$$\widehat{\text{RMSE}} = \sqrt{\Sigma n_i (X_i' - X_i)^2 / \Sigma n_i} \quad ,$$

in which X_i is the i-th raw score on the base form, X_i' is its raw score equivalent, and n_i is the number of examinees who obtained the score X_i. When there are many examinees as in this study, $\widehat{\text{RMSE}}$ is calculated more easily by the expression

$$\widehat{\text{RMSE}} = \sqrt{(m-1)^2 \text{ Var}(X) + (\widehat{\text{BIAS}})^2} \quad ,$$

in which m is the slope of the equating function $X' = mX+b$, and Var(X) is the variance of X. Note that the estimated bias, $\widehat{\text{BIAS}}$, contributes to $\widehat{\text{RMSE}}$.

Under the Tucker method, which equates on the basis of assumptions about observed scores, $\widehat{\text{BIAS}}$ and $\widehat{\text{RMSE}}$ are greater for equivalent raw scores derived from using the total examinee group than for those derived from using only

first-timers. The differences, however, are small, particularly in Chain 1 and Chain 2 (Table 2 and Table 3).

Under the Levine method, which equates on the basis of assumptions about true scores, the results do not decisively favor use of one group over the other. In Chains 1 and 2, estimated bias and root-mean-squared error under use of first-timers do consistently exceed those under use of the total group. But in Chain 3 these statistics are lower under use of first-timers. In Chain 1 and Chain 2, respective differences are slight, and they translate into small differences of raw score equivalents. For example, in Chain 2 (Table 3) a score of 160 (close to the center of the score distribution) by the Levine method equates to 160.47 under use of the total group and 160.43 under use of first-timers--a difference of only 0.04 of a raw score point on the 0 to 280 raw score scale. This instance slightly favors the use of the first-timers, whereas some of the other raw scores in Chain 1 and Chain 2 favor use of the total examinee group with the Levine method.

Equating through Chain 3 produces a much greater shift of scores than equating through Chain 1 and Chain 2. For example, a Form B score (Table 4) of 160 converts to 164.91 under the Tucker method and use of the total group. The equivalent raw score differs from the original raw score by 4.91 points. In contrast, a Form A score (Table 2) of 160 converts (Chain 1, Tucker method, total group) to 160.81, a shift of only 0.81 of a raw score point.

Repeated use of common or anchor items produced no obvious adverse effect upon the equating. A set of 49 items appeared among the anchor items of Form A, Form B, Form C, and Form D; accordingly, these items were exposed four times at 6-month intervals. Yet the two chains that involve only these forms yield much more accurate equatings than Chain 3, which incorporates Form E.

Form E marks the change to almost entirely new sets of anchor items for the Form B → Form E and Form C → Form E links.

DISCUSSION

The failure of the results to show uniformly and substantially more accurate equating under use of only first-time (nonrepeater) examinees casts doubt upon the practice of routinely excluding repeater examinees in deriving equating functions for this testing program. The overall pattern of results does appear to favor slightly the use of first-timers. But the relative accuracy of equating with first-timers versus total group also appears to depend partly on the specific equating method (Tucker Equally Reliable versus Levine Equally Reliable).

In this study, identified repeaters constituted (Table 1) from 13% to 25% of the total examinee groups. Other testing programs may have groups of examinees with greater proportions of repeaters. Further research is needed to determine whether eliminating larger proportions of repeaters produces differences greater than those found here.

The strangest results are the relatively large differences in Chain 3 between original Form B raw scores and their equated values. Chain 3 is the only one to incorporate Form E, whose links to Form C and Form D consist mostly of items previously not used as anchor items. Moreover, Form E's links to these other forms have a lower proportion of items representing the afternoon section of the day-long test. These changes accompanying Form E are likely causing the shift. If indeed they are, further investigation might reveal just how they are doing so.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 508-600.

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.), Test equating. New York: Academic Press, 9-49.

Klein, L. W., & Jarjoura, D. (in press). The importance of content balance for common-item equating with nonrandom groups. Journal of Educational Measurement.

Petersen, N. S., Cook, L. L., & Stocking, M. (1983). IRT versus conventional equating methods: A comparative study of scale stability. Journal of Educational Statistics, 8, 137-156.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating methods. In P. W. Holland and D. B. Rubin (Eds.), Test equating. New York: Academic Press, 71-135.

TABLE 1

Breakdown of Examinee Groups According to Repeater Status

	Form A		Form B		Form C		Form D		Form E	
	N	%	N	%	N	%[a]	N	%	N	%
First-Timers	10,450	78.1	22,562	85.8	11,272	74.5	11,041	72.6	23,450	83.5
Repeaters	2,725	20.4	3,385	12.9	3,484	23.0	3,807	25.0	4,115	14.6
Not Stated	207	1.5	337	1.3	367	2.4	359	2.4	530	1.9
Total Group	13,382	100.0	26,284	100.0	15,123	100.0	15,207	100.0	28,095	100.0

[a] Percentages in this column do not add to 100.0 because of rounding error.

TABLE 2
Results of Equating Base Form A to Itself
Through Equating Chain 1: Form A → Form C → Form B → Form A

Form A Raw Score	Raw Score Equivalent on Form A as Old Form: Tucker Equally Reliable Method — Total Examinee Group	Tucker Equally Reliable Method — First-Timers Only	Levine Equally Reliable Method — Total Examinee Group	Levine Equally Reliable Method — First-Timers Only
0	**-0.35**	-0.49	**-1.69**	-1.76
40	**39.94**	39.80	**38.87**	38.80
80	80.23	**80.09**	**79.43**	79.37
120	120.52	**120.37**	**119.99**	119.93
160	160.81	**160.66**	160.54	**160.50**
200	201.11	**200.95**	201.10	**201.06**
240	241.40	**241.23**	241.66	**241.63**
280	281.69	281.52	282.22	**282.19**
N	13,382	10,450	13,382	10,450
Mean	151.53	158.47	151.53	158.47
SD	35.85	34.91	35.85	34.91
$\widehat{BIAS}$[a]	0.75	0.65	0.43	0.47
$\widehat{RMSE}$[b]	0.80	0.70	0.66	0.68

Final Equating Functions

Tucker Method	Levine Method
Total Group A′ = 1.00728A - 0.35088	Total Group A′ = 1.01396A - 1.68959
First-Timers A′ = 1.00717A - 0.48616	First-Timers A′ = 1.01411A - 1.76131

Note. Within a given equating method (Tucker or Levine), the raw score equivalent closer to the original raw score appears in boldface.

Note. Means and standard deviations are those of original raw scores, not raw score equivalents.

[a] $\widehat{BIAS}$ is estimated bias or mean equating error; see page 7 for definition of this measure.

[b] $\widehat{RMSE}$ is estimated root-mean-squared error of equating; see page 7 for discussion of this measure.

TABLE 3
Results of Equating Base Form A to Itself
Through Equating Chain 2: Form A → Form D → Form B → Form A

	Raw Score Equivalent on Form A as Old Form			
	Tucker Equally Reliable Method		Levine Equally Reliable Method	
Form A Raw Score	Total Examinee Group	First-Timers Only	Total Examinee Group	First-Timers Only
0	**-0.08**	-0.29	**-1.86**	-1.95
40	40.17	**39.96**	**38.72**	38.64
80	80.41	**80.21**	**79.31**	79.24
120	120.66	**120.46**	**119.89**	119.83
160	160.91	**160.71**	160.47	**160.43**
200	201.15	**200.96**	201.05	**201.03**
240	241.40	**241.21**	241.63	**241.62**
280	281.65	**281.46**	**282.21**	282.22
N	13,382	10,450	13,382	10,450
Mean	151.53	158.47	151.53	158.47
SD	35.85	34.91	35.85	34.91
$\widehat{\text{BIAS}}$[a]	0.85	0.70	0.35	0.41
$\widehat{\text{RMSE}}$[b]	0.88	0.73	0.63	0.66

Final Equating Functions

Tucker Method	Levine Method
Total Group $A' = 1.00616A - 0.07953$	Total Group $A' = 1.01454A - 1.85730$
First-Timers $A' = 1.00623A - 0.28770$	First-Timers $A' = 1.01490A - 1.95401$

Note. Within a given equating method (Tucker or Levine), the raw score equivalent closer to the original raw score appears in boldface.

Note. Means and standard deviations are those of original raw scores, not raw score equivalents.

[a] $\widehat{\text{BIAS}}$ is estimated bias or mean equating error; see page 7 for definition of this measure.

[b] $\widehat{\text{RMSE}}$ is estimated root-mean-squared error of equating; see page 7 for discussion of this measure.

TABLE 4
Results of Equating Base Form B to Itself
Through Equating Chain 3: Form B → Form E → Form C → Form B

	Raw Score Equivalent on Form B as Old Form			
	Tucker Equally Reliable Method		Levine Equally Reliable Method	
Form B Raw Score	Total Examinee Group	First-Timers Only	Total Examinee Group	First-Timers Only
0	3.05	**2.65**	2.11	**1.80**
40	43.52	**43.04**	43.03	**42.62**
80	83.98	**83.42**	83.95	**83.43**
120	124.45	**123.81**	124.87	**124.25**
160	164.91	**164.19**	165.79	**165.06**
200	205.38	**204.58**	206.72	**205.88**
240	245.84	**244.96**	247.64	**246.70**
280	286.31	**285.34**	288.56	**287.51**
N	26,284	22,562	26,284	22,562
Mean	157.70	162.85	157.70	162.85
SD	37.32	35.94	37.32	35.94
BIAS[a]	4.88	4.22	5.74	5.12
RMSE[b]	4.90	4.23	5.81	5.17

Final Equating Functions

Tucker Method	Levine Method
Total Group $B' = 1.01163B + 3.05132$	Total Group $B' = 1.02305B + 2.10667$
First-Timers $B' = 1.00961B + 2.65313$	First-Timers $B' = 1.02039B + 1.80208$

Note. Within a given equating method (Tucker or Levine), the raw score equivalent closer to the original raw score appears in boldface.

Note. Means and standard deviations are those of original raw scores, not raw score equivalents.

[a] BIAS is estimated bias or mean equating error; see page 7 for definition of this measure.

[b] RMSE is estimated root-mean-squared error of equating; see page 7 for discussion of this measure.

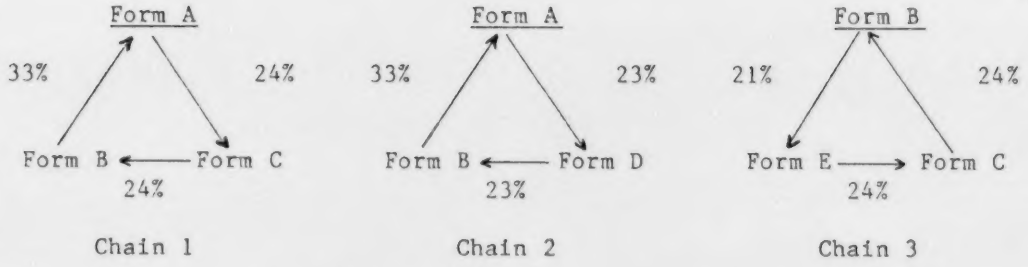

Figure 1. Chains used to compare accuracy of equating under use of first-timers only versus use of the total examinee group. Arrows indicate equating links. A percentage adjacent to an arrow gives the number of equating items as a percentage of the number of items in each of the equated forms.